The Linear Fit Programs

Matthew Marcus

October 18,2006

Updates added November, 2011

I. Introduction

These programs are for linear least–squares fitting of a XANES or EXAFS spectrum (`.e` or `.b` in my notation) to a weighted sum of a comparatively small number of reference spectra. They're designed for "physics–free" or "fingerprint" fitting. There are two programs covered in this manual, the Combination Fit program and the Linear Fit. The Combination Fit program takes a database of standard (reference) spectra and asks, "what's the best combination of three (two, one) references to fit this data as a weighted sum?". The Linear Fit program is similar, except that you tell it what combination to use, though you can turn off any of them. The advantage of the Linear Fit program, when you know what references to use, is that you can adjust E0 for the unknown or for individual references. The Combination program is good for picking among a large group of references. Despite the names, both programs perform linear least–squares fitting. Since both have some similar features and conventions, both are covered in this manual.

Important: There have been upgrades to these programs, which, in lieu of a rewrite of this manual, are described in a section at the end. That's why there are controls on the actual programs which are not shown in this manual.

II. Linear–fit program

When you invoke the program, it asks for the data to fit and then it asks 'k or E space?'. The latter is to tell it how to do E_0 adjustment, whether to allow a slope adjustment (for improving the normalization of XANES .e files), and whether to enable the Fourier transform display. It then asks for a set of reference files. You tell it you're done with entering references by hitting the Cancel button, just as in the PCA program. An example, in k–space, is shown in Figure 1a. This is the screen immediately after file entry. By default, all references are turned off, so the fit is poor as there's nothing to fit with. On turning on all references (checking the Select boxes), we get what's shown in

Figure 1b. The area of the screen around the Select boxes is shown in close-up in Figure 2. The data are shown in white, the fit in green and the residuals in red. This is a decent fit, except for one thing – the weights for the third and fifth references are negative. Since you can't have a negative amount of goethite in the sample, this must be bogus. Therefore, let's un–check the Select box for the third reference, which has the largest negative number. The fit remains OK, and the fifth reference still wants a negative weight. Let's therefore un–check its Select box. Now, we have three references with positive weights of 50.4, 53.7 and 8.1%, respectively. That the sum comes out to >100% might be related to overabsorption in the first reference, for which no correction was done. There is a Sum indicator so you don't have to do the arithmetic. That the un–checked references still show non–zero weights is irrelevant; the actual weights used in the fits are zero.

What about errors of E_0? We can shift the input data using the slider labeled Common E0. By doing this, we find a very slightly better fit with the slider at about –0.3eV. However, the improvement is very small. We can get the same effect by turning on the Slave switch (this makes all E_0 go together) and checking the Free E0 box for the last reference. This makes E_0 adjust and go to +0.2eV. The sign is opposite to that gotten by playing with the slider because the slider shifts the unknown, while the E0 boxes allow the references to move. We can test individual references by turning off the Slave switch, un–checking their respective E_0 boxes, typing 0 into the E0 indicators for those we don't want to shift, then checking the E_0 box for the one we want to test. In this case, no significant improvement is obtained.

In this example, one of the references is wanted only in a small amount (8%). Is it really needed at all? If we un–select it, the normalized sum–square (badness–of–fit) goes from 0.188 to 0.190, which isn't a huge change. On the other hand, this is unfiltered data, so one doesn't expect big changes on adding or removing one free parameter.

We can get a better look at what's going on by switching views to Fourier space. This is done by switching to the R–space tab on the graph as seen in Figure 4. You now get an un–windowed FT, which you can view in magnitude, real part, imaginary part or phase using the control (not shown in Figure 4) under the graph on the right side. Figure 4a shows the magnitude plot with three references included, with the R–scale contracted

to 5A and the lines thickened for reproduction in this manual. Figure 4b shows the same thing with the third (8%) reference turned off. The change is tiny, but visible, mostly in the second shell. It's at least plausible that the third reference adds something significant.

It should be noted that the FT here is un–windowed. It could be argued that this display should be windowed to match what is done on other programs and to make the shells stand out better. On the other hand, it could be argued that the un–windowed display is better because you're fitting all the data without weighting, so the data should be displayed that way. The obvious fix is to put in yet another slider for beta value, as in the FT program, and have it default to 0 (unwindowed). I just haven't done that yet.

Fitting of XANES data is similar to fitting EXAFS, except for two differences: there's no FT and there can be a slope adjustment. This adjustment is needed because it isn't always obvious how to do the post–edge normalization of XANES data, especially if it's taken over a short range. This adjustment consists of multiplying the input data by $1+s(E–E_{min})$, where s is a free parameter, E the energy (abscissa) and E_{min} is the minimum E in the data. Thus, if the data cover 200eV and s=0.0005, the last part of the data will be raised by 10% relative to the first. When the input data are specified to be in E–space, a slider for slope appears as shown in Figure 5a. Also, the graph won't stay in R–space mode and the Change in k–power control is grayed and disabled.

In this example, the fit is pretty good without doing anything to the slope control. In order to demonstrate the effect of slope, let's make the slope large. In order to get enough range on the slope control, we can use the Expand pushbutton to double the range of the slider a few times. Now we can crank up on slope to get Figure 5b.

If we carefully work the Slope and Common E0 controls, we can improve the fit from a normalized sum–square of 13.8e–5 to 7.44e–5, almost a two–fold difference. A close look at Figure 5a shows that the fit line with slope set to 0 is consistently above the data, whereas an adjustment of the Slope control fixes that.

There is pair of buttons for saving the fit and residual, colored in red and green to match the plots. These buttons are under the graph and will result in a filename prompt. The default extensions are `.fit` and `.res`.

Updates to linear fit program:

By default, if fitting in E-space, Allow slope? is ON.

There is a new vertical slider for overabsorption adjustment. This applies the same simple overabsorption correction to the input data that's done in `Plot, add, multiply with adjusts and overabs kludge`. There's also a button for saving the overabsorption-corrected fits. This button is under the graph, next to the other save buttons, including a new one which saves the fit fractions into a text file.

There's a button labeled Clear smallest amt. This un-selects the component with the smallest loading, and is typically used to remove those with negative loadings.

The indicator which shows the references has been expanded to allow more in view at one time and aligned with the controls for selection and E_0 control.

III. Combination fit

How did I know to use the three references I did for the above demo fit? I used the Combination fit program, which tests all combinations of a given number of references, taken from a database. A file with extension *.prm tells the program what references to use. Here is an example of such a file:

```
NbCompoMax=3
NbCompoMin=1
ref=fe2o3_mm.e
# Adjusted by comparison with a thin-film sample
ref=fe3o4_nv_overabs_adj.e
ref=ferrihydrite_2L_borch.e
# This has had a rough overabs correction
ref=Ferrihydrite_6L_borch_trunc.e
ref=Lepidocrocite_borch.e
ref=nontronite_borch.e
#ref=Fe foil XANES.e
ref=pyrite XANES.e
ref=FO0 Fe EXAFS.e
Ref = ""
```

The combination–fit program is a direct steal from A. Manceau's `fit0600` program, and the format of the `.prm` file, as well as the `prm` extension itself is designed for compatibility with `fit0600`. NbCompoMax is the maximum number of components to try. The program of course gets exponentially slower as this number goes up. NbCompoMin is not needed for this program, but is present in the older program and is in here from habit. Each reference is included by a `ref=<filename>` line, and the file ends with `ref=""`. The program is case–blind, so you can have it as `Ref=`. If simple names are given, it's assumed that the file resides in the same directory

as the `.prm` file. Otherwise, a full path name can be used. Lines preceded by # are ignored, which is handy for making notes as well as 'commenting out' references.

When we give the program the information about the input and database files, it first tries to figure out if E– or k–space is wanted, based on the input file extension. If the extension is not recognized, it asks. If it's E–space, then the program asks whether to allow a slope adjustment on each fit. This dramatically slows the program but can give a more accurate choice. No E_0 adjustment is allowed, which is part of why one often wants to use the Linear Fit program once a choice of references has been narrowed down.

Figure 6 shows what you get on entry, wherein fits of 1 component are displayed. The fit isn't very good. You can explore other 1–component fits, in descending order of goodness, by spinning the fit selector control (indicated with an arrow in Figure 6). This shows you the 1–component fits sorted in order of badness. By default, you start with the best fit showing. The two numbers displayed here are the normalized sum–squared residual and the normalized sum–abs residual $\sum|y - fit| / \sum|y|$. The sorting is always in order of the sum–square.

In Figure 7 we see what happens when you use the # of components control to allow more components into the fit. The fit gets dramatically better at 3 components. It is possible to have the best 3–component fit be worse than the best 1–component fit, paradoxical as this sounds. This is because fits with non–positive amounts are not considered, so if the best fit with components A,B and C requires a negative amount of C, then it will show up as a 2–component fit with A and B. Again, scrolling with the fit selector lets you see how well different combinations work. It is common to find that the first few combinations have similar fit qualities and differ only by the choice of a minor component. In this fit, we have allowed floating slopes, so the fit is equivalent to that in Figure 6, except that E_0 is not adjusted, and the sum–square is approximately the same as in Figure 6.

There is a control called Plot What? to the left of the graph window. This control is by default set to Fit, meaning that only the data and fit are plotted together. The other two settings are Fit and scaled components and Fit and components. The former, as shown in Figure 8a, lets you see the individual components, weighted by their amounts in the fit. Thus, the purple line shows the ferrihydrite component multiplied by

0.51, its contribution as listed in the Amounts indicator. The Fit and components selection does the same except without the multiplication. Of course, there is an FT window as in the Linear Fit program, and it's disabled when in E–space. In k–space, this window will show the scaled and unscaled components as well as the data and fit.

As in the Linear Fit program, there are buttons for saving the fit and residual.

Updates for Combination Fit program:

There's a new button, Save amounts, which writes the amounts of the fit into a tab-delimited text file. This file starts with a heading showing the filenames of all the references considered in the fit. There are three switches which modify the action of this button. One is labeled "Only list used refs?". If this switch is up, then the file will show only the names of references actually used in the fit. If the switch is down, all the reference names will be shown, which should make for very long lines of text, but is useful when doing multiple fits to the same database. If the Only list used refs? switch is down, two more switches become visible. One of these is labeled "Append to existing?". If this switch is up and if the text file selected already exists, the numbers for the current fit will be added on as a new line on the end of the file. That's useful when you're fitting many unknown to one set of references, or are trying out different fits (numbers of components, excluding some references...) to one unknown. The data line written includes the amounts of each component, including the ones not used in the fit (zero), the Sum-sq error, and the Rank, which is the setting of the fit selector. The other switch is "Suppress 0 entries?". If this is up, instead of a line like

`foo.e\t0.00000\t0.35957\t0.64902\t0.00000\t3.902e-2\t0` (`\t` = tab)

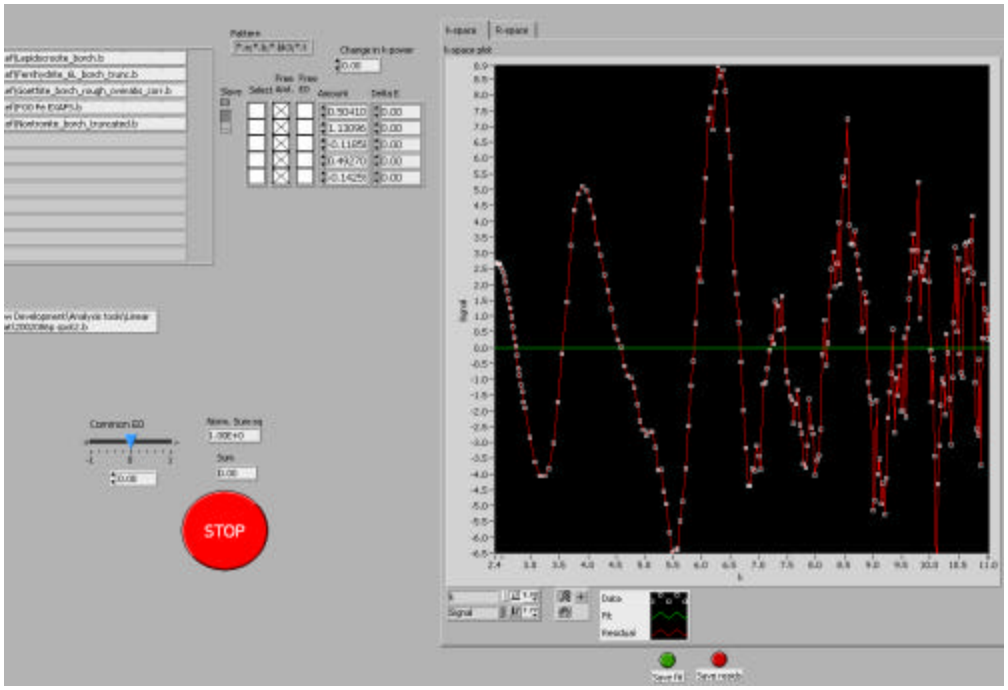
which means that unknown spectrum `foo.e` fits with the second and third references having loadings of 35.9% and 64.9%, the sum-sq is 0.03902 and this is the best fit in two components, you'll get

`foo.e\t\t0.35957\t0.64902\t\t3.902e-2\t0`.

A useful way to deal with files generated by the Save amounts button is to open it in a text editor, open an instance of Excel or equivalent, then copy-paste the contents of the

text file into the spreadsheet. The tabs will correctly translate into columns, even though that's not obvious in the text version due to the variable length of filenames.

a)



b)

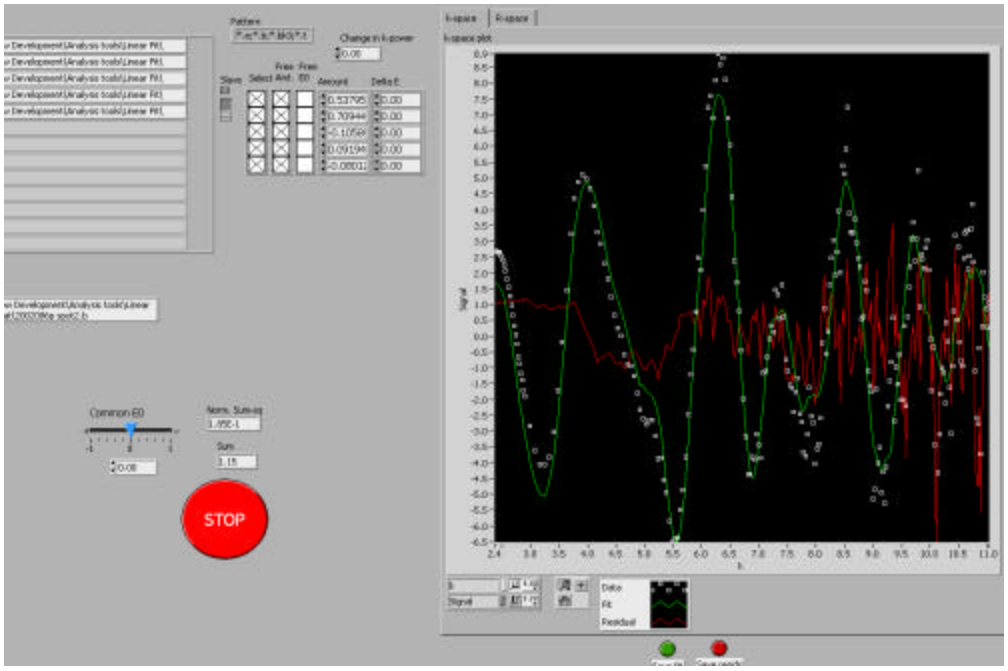


Figure 1a) Screen on entry with EXAFS data; no references are turned on; b) screen after turning on all references.

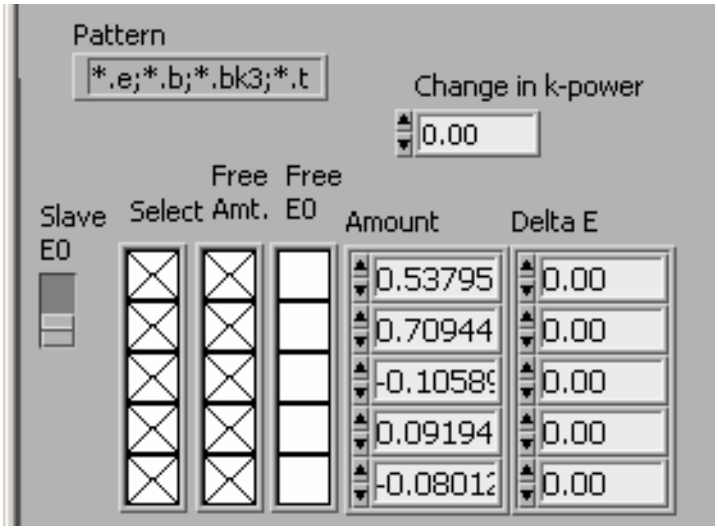


Figure 2) Closeup of the selection–box area, showing that all references are selected, none are allowed to have variable E_0, and all have floating weights. Note also that two of the references are wanted in negative amounts, which is unphysical.

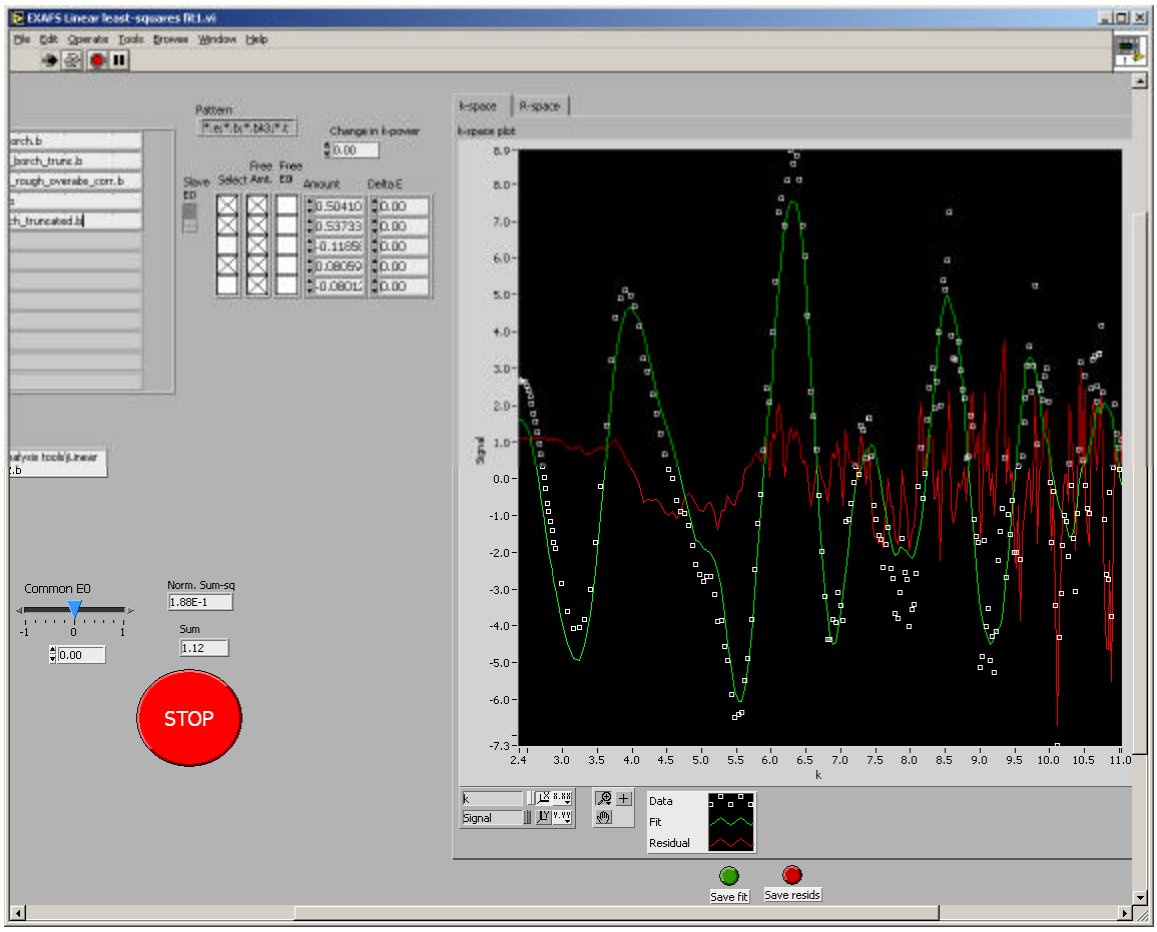


Figure 3) The screen after selecting three references. The fit is moderately good in k–space, though the residual does show some low–frequency oscillation (spline?) and a bit of higher–frequency wiggle suggesting that some of the higher shells are not perfectly represented.

a) b)

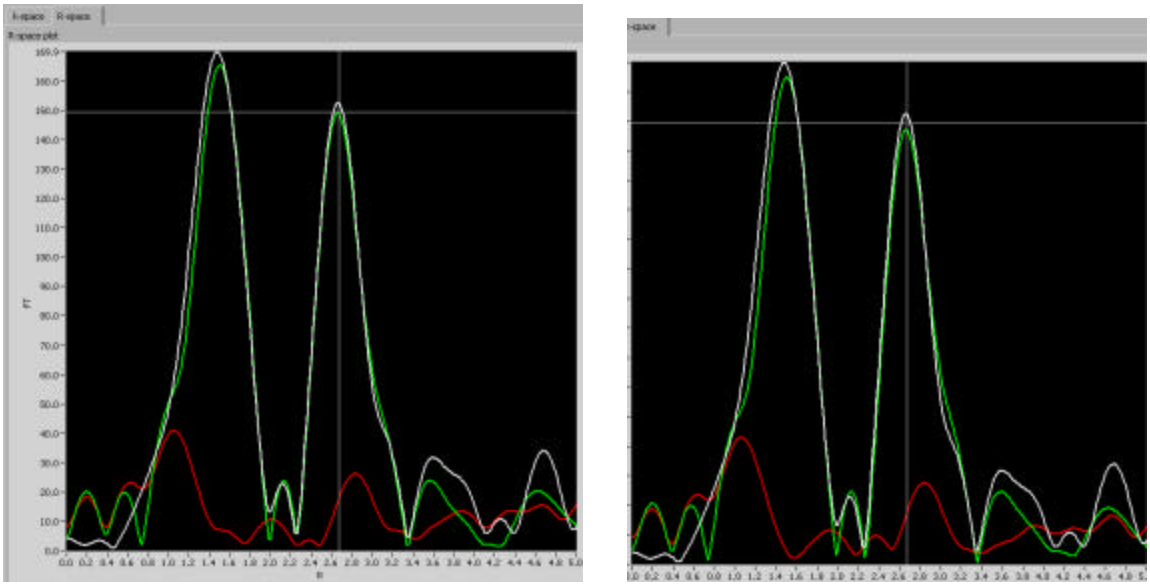
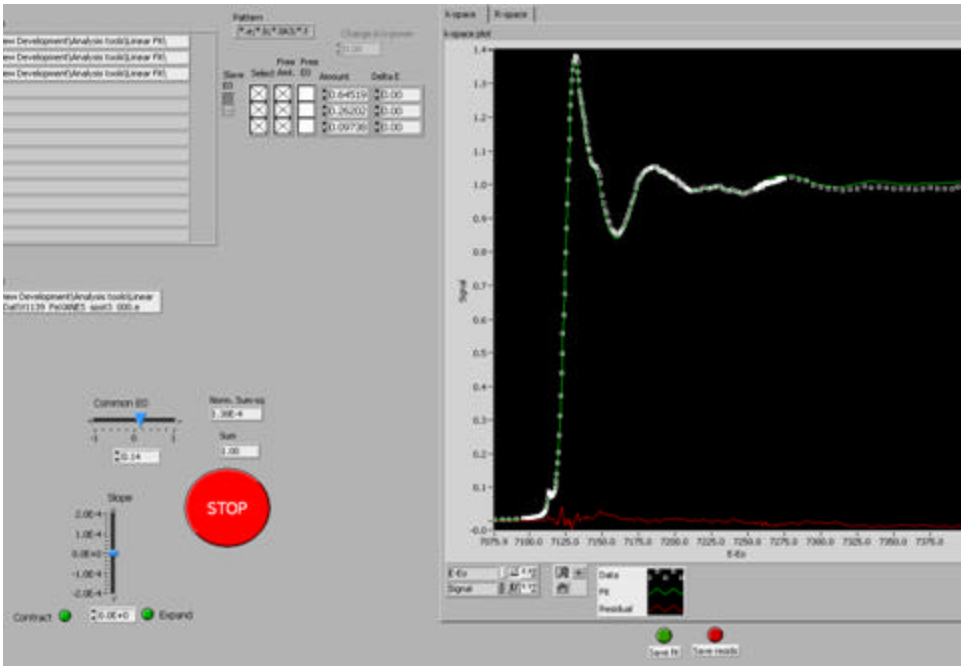


Figure 4a) The FT (magnitude) screen showing the fit with three references (a) and two (b). The second shell shows a slight change as can be seen by reference to the cursor (gray).

a)



b)

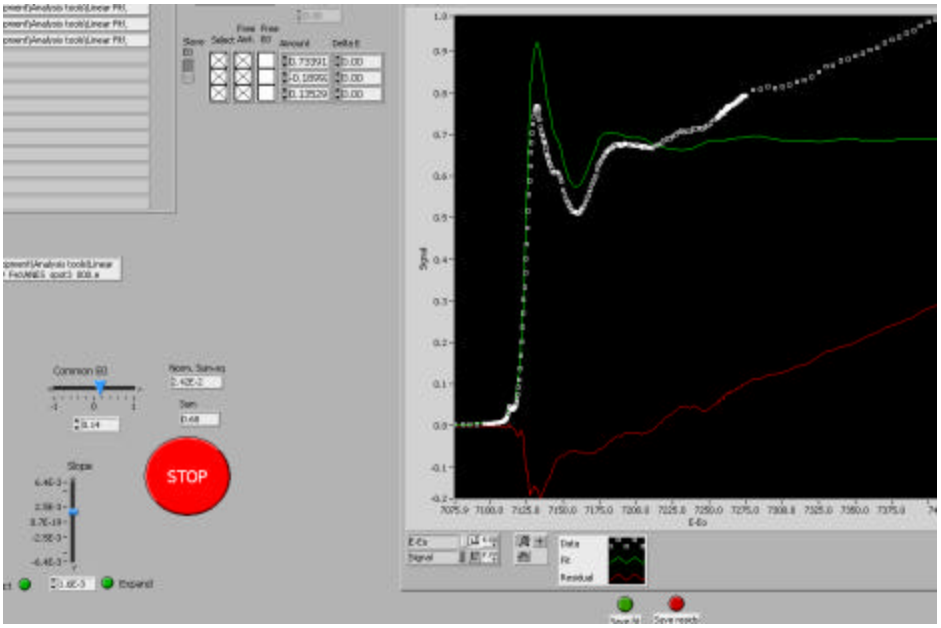


Figure 5. The main screen with the slope controls showing in the lower left. A small adjustment of E_0 has been performed, but the slope is now left at 0 (a) and a large positive value (b).

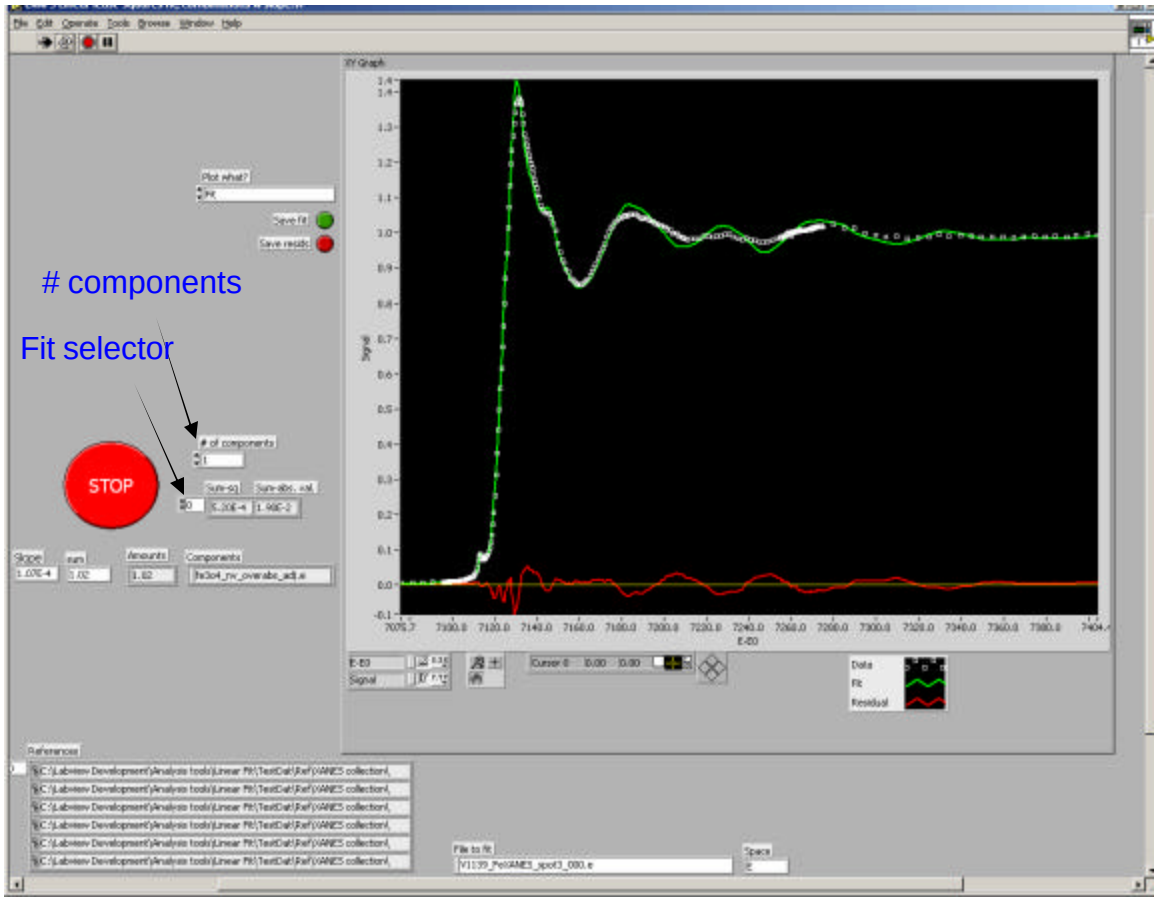


Figure 6. This shows an attempt to fit the data with one reference. Not very good, but there's none better in the database.

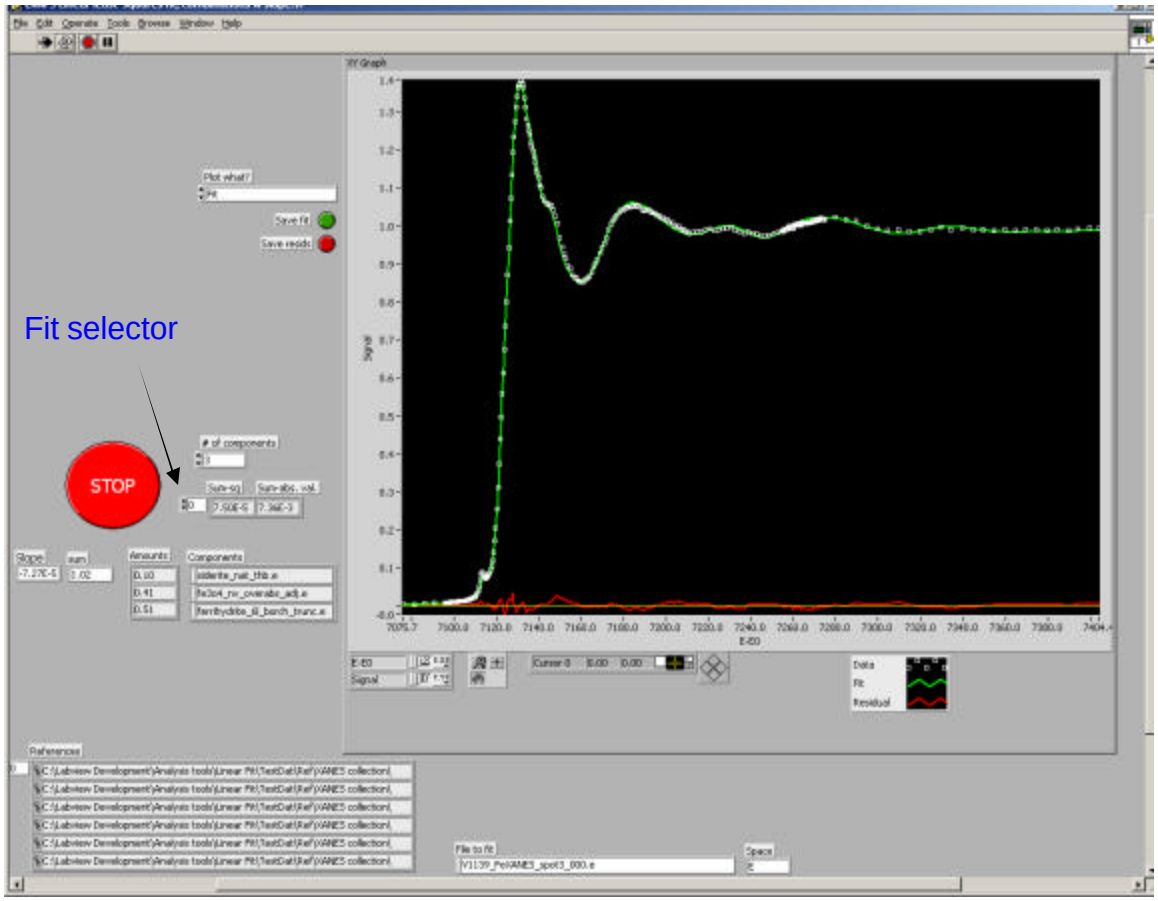
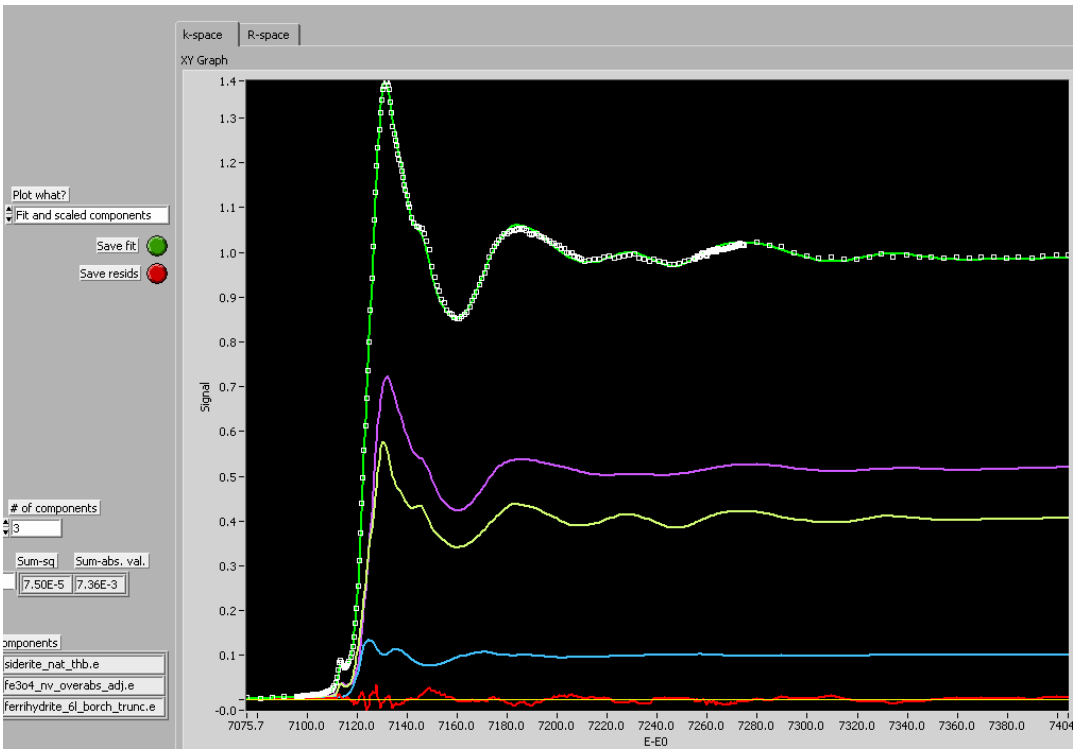


Figure 7. This shows the same data fit to three references. The data and references are the same as in Figure 5. The Fit selector is set to 0, which means that we're looking at the best fit.

a)



b)

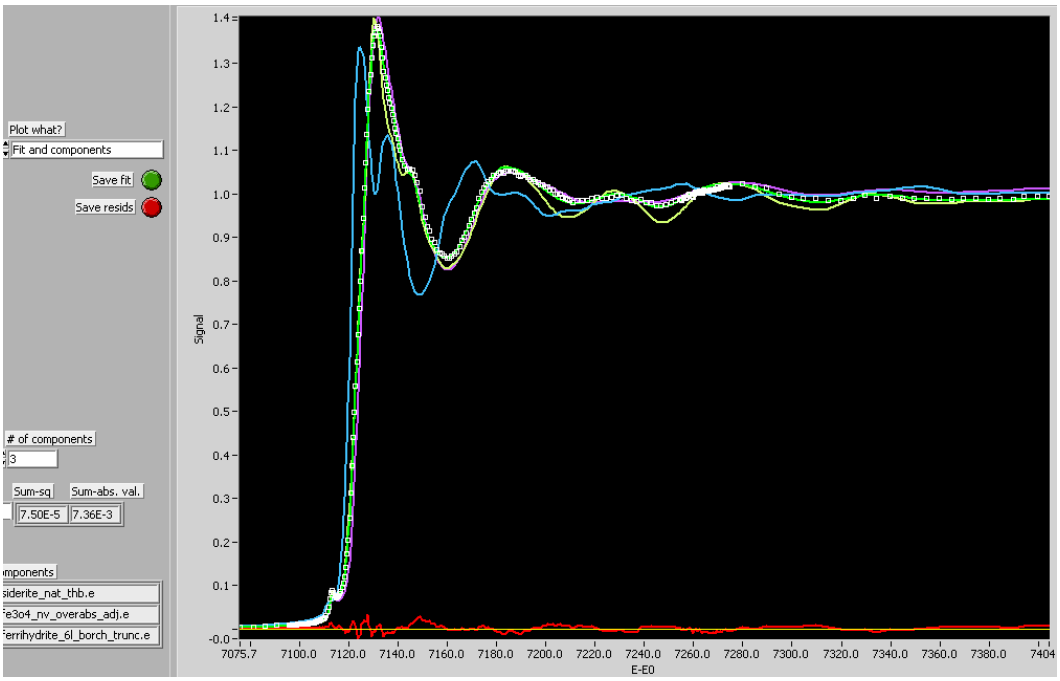


Figure 8. The same fit as Figure 7, but dissected into scaled components (a) or un–scaled components (b).